第 37 卷第 14 期
2017 年 7 月

生 态 学 报
ACTA ECOLOGICA SINICA

Vol.37,No.14
Jul.,2017

DOI: 10.5846/stxb201604210748

刘丽香,张丽云,赵芬,赵苗苗,赵海凤,邵蕊,徐明.生态环境大数据面临的机遇与挑战.生态学报,2017,37(14):4896-4904.

Liu L X, Zhang L Y, Zhao F, Zhao M M, Zhao H F, Shao R, Xu M.The opportunities and challenges presented by ecological environment big data.Acta Ecologica Sinica,2017,37(14):4896-4904.

生态环境大数据面临的机遇与挑战

刘丽香,张丽云,赵 芬,赵苗苗,赵海凤,邵 蕊,徐 明*

中国科学院地理科学与资源研究所,生态系统网络观测与模拟重点实验室,北京 100101

摘要:随着大数据时代的到来和大数据技术的迅猛发展,生态环境大数据的建设和应用已初露端倪。为了全面推进生态环境大数据的建设和应用,综述了生态环境大数据在解决生态环境问题中的机遇和优势,并分析了生态环境大数据在应用中所面临的挑战。总结和概括了大数据的概念与特征,又结合生态环境领域的特点,分析了生态环境大数据的特殊性和复杂性。重点阐述了生态环境大数据在减缓环境污染、生态退化和气候变化中的机遇,主要从数据存储、处理、分析、解释和展示等方面阐述生态环境大数据相较于传统数据的优势,通过这些优势说明生态环境大数据将有助于全面提高生态环境治理的综合决策水平。虽然生态环境大数据的应用前景广阔,但也面临着重重挑战,在数据共享和开放、应用创新、数据管理、技术创新和落地、专业人才培养和资金投入等方面还存在着许多问题和困难。在以上分析的基础上,提出了生态环境大数据未来的发展方向,包括各类生态环境数据的标准化、建设生态环境大数据存储与处理分析平台和推动国内外生态环境大数据平台的对接。

关键词:大数据;生态环境大数据;生态环境问题;环境污染;生态退化;气候变化

The opportunities and challenges presented by ecological environment big data

LIU Lixiang, ZHANG Liyun, ZHAO Fen, ZHAO Miaomiao, ZHAO Haifeng, SHAO Rui, XU Ming*

Key Laboratory of Ecosystem Network Observation and Modeling, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, *China*

Abstract: With the arrival of the era of big data and the rapid development of big data technology, the construction and application of ecological environment big data have been initiated. To comprehensively promote the construction and application of ecological environment big data, we summarize in the present paper the opportunities and advantages presented by ecological environment big data in terms of solving ecological environment problems, and analyze the challenges faced regarding the application of ecological environment big data. We review the concept and basic features of big data and present the particularity and complexity of the characteristics of ecological environment big data, taking into consideration the characteristics of the ecological environment. Subsequently, we primarily analyze the opportunities presented by ecological environment big data in mitigating environmental pollution, ecological degradation, and climate change from the aspects of data storage, data processing, data analysis, data interpretation, and presentation, and compare these with the use of traditional ecological environment data. The advantages of ecological environment big data indicate that this type of data can help us to improve the comprehensive decision-making level of ecological environment management. Although prospects for the applications of ecological environment big data in China are promising, there exist potential difficulties and challenges, such as data sharing, dataassessing, data application innovation, data management, technological innovation and launch, professional personnel training, and capital investment. Finally, we propose the

基金项目:国家重点基础研究发展规划(973)项目(2012CB417103);青海省 2013 年科技促进新农村建设计划项目(2013-N-556);四川省林业厅项目(2009-204)

收稿日期:2016-04-21; **网络出版日期**:2017-03-02

*通讯作者 Corresponding author.E-mail: mingxu@ igsnrr.ac.cn

following three priority areas concerning ecological environment big data based on our thorough review: (1) standardization of all types of ecological environment data; (2) building the storage, processing, and analysis platforms of ecological environment big data storage and processing analysis; (3) pushing forward the docking of ecological environment big data at home and abroad.

Key Words: big data; ecological environment big data; ecological environment problems; environmental pollution; ecological degradation; climate change

网络信息技术与网络通信技术的融合,极大地促进了互联网、物联网、云计算和智能传感器的快速兴起和发展,使得人类社会获得的数据信息呈爆炸式增长,大数据时代正在悄然走来[1-3]。大数据的价值主要体现在大数据的应用上,因为人们关心大数据,最终还是关心大数据的应用,关心如何从不同行业的业务需求和应用出发让大数据真正实现其所蕴含的价值,从而为人们的生产生活带来有益的改变[4-6]。整体而言,全球的大数据应用处于发展初期,中国大数据应用才刚刚起步。目前,大数据应用在各行各业的发展呈现"阶梯式"格局:互联网行业是大数据应用的领跑者,金融、零售、电信、公共管理、医疗卫生等领域正积极尝试大数据,而生态环境大数据应用则刚刚起步[4-8]。

目前,大数据在生态环境领域的应用还仅限于生态环境领域的某个方面,如环保系统,缺少跨行业跨部门的应用,不能真正体现生态环境大数据作为一个整体的优势[2, 5]。早在 20 世纪中叶,"大数据"的思想已在宏观生态学方面得到体现,例如,被称为大科学研究的国际地球物理年(1957—1958)和国际生物学计划(IBP)(1964—1974),这些研究最后演变成现在的以长期定位观测为基础的国内外生态系统研究网络,这些生态系统研究网络系统地收集和存储有关生态环境的海量观测数据[9-10]。另外,大数据在生物多样性保护和农业方面也得到了一些应用,例如,很多国家和地区已经或正在建设生物多样性信息管理系统[11];美国硅谷一家公司利用气象数据与历年农作物产量进行关联分析,预测各地农场来年产量和适宜种植品种,并以个性化保险服务向农户出售[4]。在环境领域,大数据在美国环境污染防治管理中得到了初步的应用[12]。2016 年 3 月,我国环保部发布了《生态环境大数据建设总体方案》[13],为环保系统开展生态环境大数据建设提供了强有力的政策支持和技术框架,这也意味着大数据在我国环境领域的应用才刚刚起步。鉴于以上大数据在生态环境领域的应用现状,本文阐述了生态环境大数据的独有特点,梳理了大数据在解决生态环境问题中的优势和机遇,总结了生态环境大数据建设所面临的挑战,并提出了生态环境大数据未来的发展方向,为生态环境大数据的发展和大规模应用提供依据。

1　大数据概述

虽然各国都在积极准备迎接大数据时代的到来,但国内外关于大数据的定义、内涵和标准还没有达成统一认识。根据大家对大数据定义有着不同的侧重点,可以将其分为三类:第一类定义主要突出"大"[4-6],例如麦肯锡、IDC、亚马逊、维基百科等给出的定义,"大"只是大数据的重要标志之一,但并不是全部[6]。第二类定义主要是突出其"作用",认为大数据是在多样或者大量数据中,迅速获取信息的能力,与第一类只从数据本身出发不同,该类定义强调大数据的功能和作用[6]。第三类定义主要突出其"价值观和方法论",认为大数据是用崭新的思维和技术对海量数据进行整合分析,从中发现新的知识和价值,带来"大知识"、"大科技"、"大利润"和"大发展"[7]。但随着全球数据的飞速增长,除了包含传统的结构化数据,还产生大量非结构化数据和半结构化数据,这就需要大量处理技术来处理这些不同结构的数据,并将它们应用在实践中[4-7]。因此,大数据不仅仅包含海量数据,还应包括各种大数据技术的集合和大数据在各领域的应用[6]。综合以上信息,我们认为大数据是为决策问题提供服务的大数据集、大数据技术和大数据应用的总称。

目前对大数据普遍认可的是其具有以下"5V"特点[4, 14]。第一,数据量巨大。通过各种设备产生的海量

数据,规模庞大,数据量从 TB 级别跳跃到 PB 级别[4-7]。第二,数据种类繁多。数据来源种类多样化,不仅包括传统结构化数据,还包括各种非结构化数据和半结构化数据,而且非结构化数据所占比例越来越高[4-8]。第三,大数据的“快”,包括数据产生快和具备快速实时的数据处理能力两个层面。第一层面是数据产生的快。目前有的数据是爆发式产生[4,14-15],例如,欧洲核子研究中心的大型强子对撞机在工作状态下每秒产生 PB 级的数据;有的数据是涓涓细流式产生,但是由于用户众多,短时间内产生的数据量依然非常庞大,例如,点击流、日志、射频识别数据、GPS(全球定位系统)位置信息[15]。第二层面是对数据快速、实时处理的能力高。大数据技术通过发展不同于传统的快速处理的算法,对海量动态数据进行处理分析,使它们变为可使用的有价值数据。因此,大数据对实时处理有着较高的要求,数据的处理效率就决定着获得信息的能力[4,14]。第四,数据价值密度低、应用价值高。大量不同数据集组成大数据集,这些数据集的价值密度的高低与数据集总量的大小成反比。在大数据应用中,数据量大的数据并不一定有很大的价值,不能被及时有效处理分析的数据也没有很大的应用价值[4-8]。第五,真实性低。随着社交数据、企业内容、交易与应用数据等新数据源的兴起,我们能获得的数据源逐渐多样化,这使得获得的数据中有些具有模糊性[16]。真实性将促使人们利用数据融合和先进的数学方法进一步提升数据的质量,从而创造更高价值。例如,社交网络中的视频、语音、日志等获得的原始数据真实性差,需要我们对其过滤和处理才能挑出有用的数据。

2 生态环境大数据特点

大数据在解决生态环境问题时形成了生态环境大数据独一无二的特征。第一,生态环境大数据具有“空天地一体”的巨大数据量。从数据规模来看,生态环境数据体量大,数据量也已从 TB 级别跃升到 PB 级别。随着各类传感器、RFID 技术、卫星遥感、雷达和视频感知等技术的发展,数据不仅来源于传统人工监测数据,还包括航空、航天和地面数据,他们一起产生了海量生态环境数据。例如,2011 年世界气象中心就已经积累了 229TB 的数据[6];我国林业、交通、气象和环保等数据量级也都达到了 PB 级别,而且还在以每年数百个 TB 的速度在增加[17-19]。第二,生态环境大数据的类型、来源和格式具有复杂多样性[20]。从数据种类来看,生态环境数据类型多,数据来源渠道广,结构复杂。首先,生态环境数据来自于气象、水利、国土、农业、林业、交通、社会经济等不同部门的各种数据;其次,大数据技术的发展使得生态环境领域的研究不再局限于传统结构化数据类型,使得各种半结构化和非结构化数据(文本、项目报告、照片、影像、声音、视频等)的应用与分析成为可能,例如,一段历史电影视频中关于气候的描述;公众移动手机拍摄的关于植物类别的图片等;再次,来源于不同部门的同一种数据其格式多样,目前无统一的标准规范,使得难以整合和合并不同部门之间的同类数据。第三,生态环境大数据需要动态新数据和历史数据相结合处理[13]。从数据处理速度来看,由于生态系统结构与功能的动态变化而引起的生态环境数据具有强烈的时空异质性,生态环境数据多表现为流式数据特征,实时连续观测尤为重要。只有实时处理分析这些动态新数据,并与已有历史数据结合起来分析,才能挖掘出有用信息,为解决有关生态环境问题提供科学决策。第四,生态环境大数据具有很高的应用价值。从数据价值来看,生态环境大数据无疑具有巨大的潜在应用价值,利用大数据技术从海量数据中挖掘出最有用的信息,把低价值数据转换为高价值数据,最终,高价值大数据为解决各种生态环境问题提供科学依据,从而改善人类生存环境和提高人们生活质量;第五,生态环境大数据具有很高的不确定性。从数据真实性来看,虽然应用于生态环境领域的各种传感器监测精度都很高,正是因为这一点仪器往往会顺带记录大量的周边环境数据,而我们感兴趣的数据可能会埋没在大量数据中,因此,为了确保数据的精准度,需要利用大数据技术从海量数据中去伪存真,获取真实数据[9]。

3 大数据在解决生态环境问题中的优势和机遇

20 世纪后半叶以来,随着经济的发展,全球生态环境问题日趋严重。目前全球生态环境问题突出表现在环境污染、气候变化、土地退化、森林锐减、生物多样性丧失以及水资源枯竭等方面[21]。这些问题往往涉及尺

度大、过程复杂、驱动因素众多,解决起来难度大。随着大数据时代的到来,大数据为各种生态环境问题的解决提供了新的机遇。

3.1　大数据在解决环境污染中的优势

随着工业化、城市化、化学农业和机动化的高速发展,全球环境污染日益加剧,以大气污染、水污染和土壤污染为主的三大污染引起的食品安全和人类健康问题严峻,直接威胁到人类的生命[22]。如何有效的治理这些污染,是各国政府及学者迫切需要解决的难题。然而,这些污染的产生受到多方面的影响,治理起来相当困难。首先,环境污染涉及的过程复杂,包括污染物排放的生物过程、污染物在承载体(大气、水和土壤)中的物理和化学过程;其次,污染成因很多,主要包括工业三废(废水、废气和废渣)、农业污染(肥料、农药和农膜)、机动车尾气排放、生活垃圾以及木材和煤等燃料燃烧;最后,影响污染因素多,因素之间存在相互重叠和交叉作用。因此,仅靠传统单因素单独治理污染不能解决根本问题,这就需要通过利用云计算、多元数据同化、多尺度数据耦合、时空分配和化学物种分配等大数据技术对各种环境污染及其相关的数据进行多因素融合分析,及时准确地发现各种污染的根源,分析不同污染过程中污染物的演变规律,了解各种主要污染物的"前世今生",全面地获得污染物的变化规律和传输过程,通过这些信息来区分环境污染的轻重缓急,统筹规划治理方案,分步推进污染治理,既要综合治理也要重点突破[5,12]。

另一方面,环境污染对人类影响具有滞后性,污染发生时很难感知和预料,但这些影响一旦产生就表示已经发展到相当严重的地步。因此,除了增强污染事后治理, 还需加强污染事前预防。当前环境污染很大程度上还只限于治理,很少采取预防措施,更缺少对重大环境污染事件的预报预测。目前,我国环境污染的预测预报主要是通过各种数据建立统计模型,但这些模型的参数缺少优化,预报预测准确性低[12]。例如,我国已经开发了一些污染物扩散预测模型,可由于缺乏这些污染物长期实时数据,不能对模型参数优化,使得预报预测的准确性低。大数据时代的到来,为提高我国环境污染预报预测带来了机遇。随着云计算、机器学习和人工智能等技术的不断发展,使得建立基于认知计算的高精度环境污染预报系统成为可能。环保部门积累的环境污染应急管控经验可以加入认知计算系统,使得应急管控变为常态管理,例如,可以将专家经验加入认知计算系统中。认知计算整合优化各类模型,包括物理化学过程、气象、交通和社交等,它们再通过海量数据进行交叉验证,该算法使模型、数据和专家经验以自动训练、自我思考和自我学习的方式不断积累,为可靠追溯污染源、高精准预报预测、精细预防和治理等决策提供科学支撑[12]。

3.2　大数据在改善生态退化中的优势

随着全球人口数量的增长和社会经济的发展,生态系统退化越来越严重,已经成为全球严重的生态环境问题之一。当前全球生态退化主要表现在森林面积减少、土地退化、生物多样性降低、水资源短缺等方面,这些退化引起了全球森林资源、水资源和土地资源的减少。生态退化除了造成巨大经济损失,还严重威胁到人类健康和生命安全[21]。

首先,引起生态退化因素较多,主要包括乱砍滥伐、过度农垦、陡坡开垦、生境丧失、生物资源过度开发、水环境遭破坏、外来物种入侵、海洋的过度捕捞以及环境污染等[32-34]。以上因素相互交织,协同作用,致使一种生态退化类型可能是另一种退化的原因,例如,森林面积减少可引起土地退化、生物多样化减少、水资源短缺加重。另外,生态退化是一个复杂和综合的动态过程,它涉及跨领域、跨学科、跨部门的各种生态环境数据,又与社会、经济、文化和政策等领域密切相关;同时涉及土壤、农学、生态、环境和生物等学科的知识。过去几十年,虽然各国政府也采取了一些措施治理生态退化,但由于生态退化所涉数据来源多样、分布广泛,内容庞杂、涉及部门众多,而传统技术不能系统地整理和分析这些数据集,也不能完全提纯出数据背后的有价值信息,或者由于技术落后提炼出的信息为错误的,以这些错误的科学数据信息作为理论指导,使得政府的经济政策和防治决策对生态退化没用,甚至失误[35]。目前,随着大数据的蓬勃发展,人们可以利用传感器技术和无线通信技术在数据获取方面的优势,系统地收集、整理和存储各种与生态退化相关的数据,包括地面监测数据、遥感影像数据、社会经济数据、科学研究数据、互联网以网站、论坛、微博等方式发布的有关资源环境的相关信

息,实现了生态环境数据的整合和充分利用,为生态系统的资源管理、生态环境的动态监测和生态环境评价提供多样化、专业化和智能化的数据服务;利用分布式数据库、云计算、人工智能、认知计算等技术在大数据处理方面的优势,并结合大数据各种算法库、模型库和知识库分析这些不同结构的数据,实现数据与模型的融合,挖掘隐藏在海量数据背后的各种信息[29-30],通过这些信息既可以分析各种生态系统退化的过程和规律,也可以为决策者提供360度的数据信息,为治理和预防生态退化提供正确的科学决策。例如,使用Hadoop的分布式文件系统(HDFS)和分布式数据库(MapReduce)对生态环境大数据进行批量处理;利用决策树、贝叶斯、K-Means、岭回归模型、逻辑斯蒂模型、线性回归模型、认知算法、关联规则的Apriori算法等各种模型和算法对海量数据进行深度挖掘和关联分析,通过各种数据的碰撞产生出有价值的信息。

3.3 大数据在减缓气候变化中的优势

近百年来,由于气候自然波动和人类活动引起的温室效应,地球气候正经历一次以全球变暖为主要特征的显著变化。全球变暖导致了极端气候出现频率增加、厄尔尼诺现象加剧且影响范围变大、冰川萎缩、内陆冻土加剧融化、沙漠化加剧、海平面上升和海水倒灌、水资源短缺加重、湿地面积减少和生物多样性下降。例如,在2001—2010年,全球冰川平均质量年下降速度为0.54 m(相当于水当量)[36]。全球变暖除了引起全球气候变化,还对农业、生态环境和人体健康产生了巨大的影响。大气中温室气体浓度增加引起了大气温室效应增强,并最终导致了全球气候变暖,温室气体主要包括CO_2、CH_4和N_2O。为了减缓和预测全球变暖的速度,政府间气候变化专门委员会(IPCC)编制了各种温室气体的排放源和吸收汇的全球清单,并预测了未来全球温度的变化;各个国家也都根据本国实际拥有数据情况编制国家温室气体清单。但目前这些温室气体清单还都不是实时清单,都是温室气体排放和吸收的总量。这主要是因为缺少温室气体的实时监测数据和缺少处理海量数据的技术。在大数据时代,网络信息技术和无线通信技术的融合,极大地促进了各种智能传感器的快速兴起和发展,使我们可以获得温室气体、气候等大量实时监测数据和与之相关的非结构化数据;基于云计算环境下,分布式数据存储技术与传统的关系型数据库相结合可以解决海量数据的存储和管理,例如,Hbase、Redis和Key-Value等大数据存储技术[37-40];同理,这些海量温室气体、气候和其他相关数据的处理分析也需要各种模型和算法,但对于编制实时温室气体清单来说,最关键技术是怎样在线和离线相结合对海量数据进行分析?离线静态数据的大数据处理形式是批量处理,Hadoop是典型的批量数据处理系统[29-30];在线数据的大数据处理形式包括实时流式处理和实时交互计算两种,流式数据处理系统如Storm、Scribe和Flume等,交互式数据处理系统如Spark和Dremel。另外,利用大数据技术融合温室气体数据和气候模型,预测未来温度的变化速度,例如,人工智能和认知算法等大数据技术。通过编制实时温室气体清单和预测未来温度变化幅度,可以为制定减排措施提供科学依据,同时也为人们的生活带来方便。可以发现,生态环境问题彼此相互联系,相互影响,相互制约。因此,治理和预防需要对区域甚至全球的生态环境情况进行全面分析,找到关键问题与关键区域,制定不同的解决方案与对策,通过对比分析找到最优解决途径。利用大数据在数据采集、数据存储、数据分析,以及数据解释和展示等方面的优势,有利于揭示生态环境问题的本质,并分析其背后的驱动因素及相互作用机制。在数据采集方面,通过建立高密度、全区域和多方位的监测网络体系[8,12],配合文本、图片、XML、HTML、各类报表、图像和音频/视频信息等与生态环境相关的非结构化数据和半结构化数据的采集,共同形成生态环境大数据集。在数据存储方面,NoSQL(Not only SQL)数据存储包括分布式文件系统和分布式数据库二种类型[26]。通过与大数据的NoSQL数据存储管理技术相结合,克服传统关系型数据库经常由于采用分片技术而出现的存储空间不够、数据加载缓慢和排队加载等问题[23-25]。在数据分析方面,我国生态环境相关的数据大多是数据集成,供客户端自行下载分析;而大数据分析却能将统计分析、深度挖掘、机器学习和智能算法与云计算技术结合起来[27-29],对空气、土壤、水文、生物多样性、气候、人口和社会经济等数据进行关联性分析,这些分析结果可为管理者的决策提供科学支持。除此之外,在数据解释和展示上,传统数据显示方式是用文本形式下载输出,而大数据却可以给用户提供可视化结果分析[29-30]。由此可见,只有大数据时代我们才能够真正实现复杂生态环境问题的定量评估和精准决策,为加快我国生态文明建设和促进生态环保事业

的发展提供科学依据和有效对策。

4　生态环境大数据面临的挑战

虽然大数据为解决各种生态环境问题提供了新的机遇,然而生态环境大数据的大规模应用才刚刚起步。生态环境大数据的真正实施在数据开放和共享、大数据处理技术、资金投入、专业人才、应用创新和数据管理等方面还面临着诸多挑战。

4.1　缺乏数据共享

生态环境大数据需要整合和集成政府多部门和社会多来源的数据(例如个人和企业等),只有不同类型的生态环境大数据相互连接、碰撞和共享,才能释放生态环境大数据的价值。因此,要想挖掘隐藏在生态环境大数据背后的潜在价值,实现数据共享是关键,也是解决生态环境问题的前提和基础。然而,实现数据共享还面临巨大挑战。首先,我国生态环境大数据包括气象、水利、生态、国土、农业、林业、交通、社会经济等其他部门的大数据,涉及多领域、多部门和多源数据,虽然目前这些部门已经建立了自己的数据平台,但这些平台之间互不连通,只是一个个的“数据孤岛”[8, 12]。大部分数据只是公开,而非开放,即数据只是发布和公开,而无法下载和利用数据[12],仅限于“看”,而无法真正去“用”,很多生态环境数据还在档案柜里“睡大觉”。其次,数据没有规范化,数据存储格式不一样,即使在同一个行业,数据也是“一人一个模样”,形成了“拥有者不一定觉得有用,看得懂、用得着的不一定能拥有”的局面。我国至今还有大量与生态环境相关的历史资料还不是电子形式,由于缺乏有效的数字化技术和手段,早期积累的很多纸质档案资料面临破损与消失的风险,这些宝贵档案资料的数字化也是一个较大的挑战。另外,数据开放严重不足,主要表现在数据开放总量偏低,可机读性差,大多为静态数据,且集中在经济发达、政府信息化基础和 IT 产业发展好的城市。最后,生态环境数据的整合和脱敏也是一项重大挑战,因为开放数据即任何人都能自由下载和利用机器可读的数据格式,所以哪些数据可以公开,哪些数据敏感,需要脱敏等等,这些都是需要耗费巨大人力物力的工作。

4.2　缺乏技术创新和落地

在数据来源方面,生态环境大数据来源多种多样,既包括各种“空天地”的监测和调查数据,也包含各种影像、声音和视频等非结构化数据,这些庞大的数据杂乱无章、参差不齐,如何将这些多源异构数据转换成合适的格式和类型,并在存储和处理之前对采集的数据进行去粗取精,并保留原有数据的语义以便后面分析,是生态环境大数据面对的一个技术挑战。目前常用的是通过数据清洗和整理技术对其填补数据残缺,纠正数据错误,去除数据冗余,将所需的数据抽取出来进行有效集成,并将数据转换成要求的格式,从而达到数据类型统一、数据格式一致、数据信息精练和数据存储集中等要求[29-30, 41]。例如,LSI 公司开发了一款多核处理器可对数据进行实时分类,降低网络流量。在数据存储方面,当前生态环境大数据由于各种移动终端和网络的视频、文本、图片、照片等非结构性数据流正在爆发性增长,未来存储技术的效率对于提高大数据的价值至关重要,包括存储的成本和性能。相比于传统的物理机器存储(包括单机文件和网络文件系统),适用于生态环境大数据的分布式存储系统提高数据的冗余性、可扩展性、容错能力、低成本和并发读写能力。例如,LSI 的闪存技术可以大大提升数据的应用速度。因此,需要不断研发进行存储技术创新,将操作便捷性的关系型数据库和灵活性的非关系型数据库融合,是未来技术创新的发展目标。在数据分析方面,目前 Google 的 MapReduce 系统、Yahoo 的 S4 系统、Twitter 的 Storm 系统、Pregel 系统等分别从离线批量计算、实时计算、图数据处理[37-39],都是针对不同的计算场景建立了不同的计算平台,管理运营成本很高,所以研发适合多种计算模型的通用架构是生态环境大数据建设和发展的急切需求。另外,数据分析已经从传统的通过先验知识人工建立数学模型到建立人工智能系统,通过人工智能和机器学习技术分析生态环境大数据是未来解决生态环境问题的关键手段。但对于他们的深度学习还需要大量工程和理论问题[42-44],例如,基于深度神经网络的机器学习,其模型的迁移适应能力以及大规模神经网络的工程实现。众所周知,工具、开源以及框架设施是大数据技术发展的方向,因此,当前大数据的技术创新形成了“互联网公司原创——开源扩散——扩散制造商产品

化——其他企业使用”的产业链格局。不过,要想实现生态环境大数据的技术和应用一体化发展,企业和政府部门必须抛弃“拿来主义”态度,只有加强对技术开源社区的贡献,才能加强对技术的深入理解,也才能更好的发挥大数据在生态环境领域的应用[41]。同时,还要加强管理制度配套和工作人员能力提升等方面,实现技术落地[8]。

4.3 资金投入不足

目前,国内外对生态环境大数据的资金投入不足。缺乏大数据重大示范项目,大部分国家缺乏生态环境监测设备、计算机资源和数据资源等基础设施的投入,包括网络服务器、数据处理和存储系统、数据仓库系统、云计算平台等。同时也缺乏对生态环境大数据拓展融资渠道, 缺少地方政府、工商企业和有实力、有需求的生产经营主体参与大数据融资。还没有成熟的大数据产业推广模式。

4.4 缺乏大数据专业人才

大数据时代的到来,对各国现有教育体系提出了全新的挑战。大数据时代需要大量的复合型人才,尤其是生态环境大数据涉及的学科众多,既需要计算机、通讯等工程技术,也需要数学、统计、人工智能等模型技术,更需要生态、环境、气象、水文、土壤等专业知识。当前许多地区的教育体系不符合未来生态环境大数据发展的战略需要,尤其是现有的高等教育体系学科分类明确,独立性比较强,缺乏学科之间的交叉融合。很多地方还没有开设大数据相关的专业和课程,也缺少大数据环境监测、生态信息学和环境信息学等方面人才培养。

4.5 应用活力不足

我国生态环境大数据的创新应用还很有限,大数据的威力远远未能发挥出来,政府综合运用生态环境大数据的能力较低,没有形成成熟的生态环境大数据产业链和有影响力的数据企业。生态环境大数据在气象、水利、国土、农业、林业、交通、社会经济等各部门的应用才刚刚起步,跨领域的应用寥寥无几。如何促进大数据在生态环境领域中的应用创新,使大数据真正成为提高生态环境监管能力现代化的有力手段,是目前世界各国正在探索的课题。

4.6 缺乏数据管理

2015 年 9 月 5 日,国务院公开发布《国务院关于印发促进大数据发展行动纲要的通知》(以下简称《纲要》)。《纲要》系统部署了大数据各项工作,并指出大数据已成为提升政府治理能力的新途径。2016 年 3 月,环保部刚刚发布了《生态环境大数据建设总体方案》,为环保系统开展生态环境大数据建设提供了强有力的政策支持和技术框架。在大数据时代,我国政府严重缺乏对数据的管理,同时在利用大数据治理生态环境问题的方式上也面临严峻挑战。

首先,政府生态环境领域职能部门缺乏“大数据”思维和意识。我国已经数字化的生态环境数据资源数量和质量都表现出“双低”状态,例如,很多纸质档案资料面临破损与消失的风险,如气象资料。有些政府部门不知道自己有什么数据,自己甚至没有“数据清单”。另外,生态环境大数据目前还没有形成统一标准的数据格式,地方和各个系统都在制定自己的数据标准,目前急需对数据格式进行统一的标准规范,这是实现数据共享和开放的关键[8,12,20]。

其次,政府的现代管理理念和运作方式不适应“大数据”管理决策的要求。生态环境大数据开发的根本目的是以数据分析为基础,帮助政府在解决生态环境问题的过程中作出明智的决策。因此,要改善我们政府的管理模式,需要管理方式和整体结构与大数据技术工具相适配[8,12]。例如,在应急管理的事前准备、事中响应和事后救援与恢复的每一阶段都可以引入大数据的应用,每个阶段对大数据的应用程度也会因其需要应对内容的不同而有所差别。如果各个部门不能改变管理模式和协同配合,常造成人为的损害。例如,最近我国南方遭遇的台风和强降雨事件,如果人们利用大数据的思维去管理,可以通过收集地面气象站和卫星的温度、风速和降雨量的小时数据,对台风和降雨进行预测时空分布,可以事前疏散大众,挽救国家和人民财产及生命。

最后,生态环境大数据面临严重安全隐患。大数据的安全主要包括大数据自身安全和大数据技术安

全[45],大数据自身安全指在数据采集、存储、挖掘、分析和应用过程中的安全,在这些计算和存储过程中由于黑客外部网络攻击和人为操作不当造成数据信息泄露,外部攻击包括对静态数据和动态数据的数据传输攻击、数据内容攻击、数据管理和网络物理攻击[46-49]。例如,很多野外生态环境监测的海量数据需要网络传输,这就加大了网络攻击的风险,如果涉及到军用的一些生态环境数据,本来人们可以国内共享,但如果被黑客获得这些数据,就可能推测到我国军方的一些信息,后果不堪设想。大数据技术安全是利用大数据技术解决信息系统安全的问题[45-48],即黑客利用大数据技术对生态大数据进行攻击,轻松获得很多涉及国家机密和比较敏感的生态环境领域的数据。随着云计算技术的发展,数据在云端的存储存在严重的安全隐患。例如,美国"棱镜门"事件,美国政府就是通过云计算和大数据技术收集大量数据也包括各国生态环境敏感数据。因此,我国未来应加强生态环境大数据安全技术研发、生态环境大数据信息安全体系的建设和管理等方面。

5　结论与展望

本文对生态环境大数据在解决生态环境问题中所面临的优势和挑战进行了系统的梳理和概括总结。与传统生态环境数据库相比,生态环境大数据不仅仅是各类生态环境数据的集成,它是对各种生态环境数据进行了深入分析并与其他相关数据进行关联分析后的数据产品,同时生态环境大数据还能对未来生态环境存在的重大风险进行预测预报,并给管理者提供科学的决策。在数据获得方面,除了政府部门的数据外,生态环境大数据也包含各类市场主体、社会组织、科研教育机构等各类团体与个人所拥有的大量与生态环境相关的数据。在数据存储和处理方面,利用各种大数据技术与传统技术相结合处理生态环境的静态、实时和图的海量数据。在数据分析和挖掘方面,借助算法库、模型库、云计算、人工智能、知识库对生态环境大数据进行深度挖掘、认知计算、关联分析、趋势分析、空间分析等各类信息挖掘,实现数据与模型的融合,开发新的数据产品提升大数据的应用价值。在数据解释上,生态环境大数据可以提供给用户可视化大数据挖掘展示。今后要不断加强大数据技术研发、加强资金投入、加强复合型人才培养、加强数据开放共享和加强生态环境大数据管理等方面,最终实现生态环境决策管理定量化、精细化,生态环境信息服务多样化、专业化和智能化,为我国社会经济可持续发展和生态文明建设奠定基础。

此外,鉴于大数据在解决生态环境问题中面临的机遇和挑战,借助云计算、人工智能及模型模拟等大数据分析技术,生态环境大数据未来迫切需要开展以下研究。(1)对各种生态环境数据进行数据标准化处理。由多个部门组成专门机构调研决定数据的技术规范与标准,搜集、整理、加工已有各个部门历史生态环境数据,实现各部门生态环境数据资料的集成。(2)依托现代数据存储与处理分析技术,构建生态环境大数据存储与处理分析平台,实现生态环境大数据的查询、更新和维护、备份等功能,在此基础上,对生态环境数据进行集成分析和信息提取。(3)推动生态环境大数据与国内外同类数据平台的对接。推动生态环境大数据与农业农村大数据、工业和新兴产业大数据、以及医疗健康和交通旅游服务大数据等大数据平台的对接,探索各相关部门数据融合和协同创新应用,实现现代农业可持续发展、减少工业污染及碳排放、流行性疾病的预防以及重点景区生态环境保护、风险预警等;加强国际交流,使我国生态环境大数据分析技术与国际接轨;为解决跨国界跨区域的全球性生态环境问题提供科学依据。

参考文献(References):

[1] Nature. Big Data. [2014-08-23]. http://www.nature.com/news/specials/bigdata/index.html.

[2] Jonathan T O, Gerald A M. Special online collection: dealing with data. Science, 2011, 331(6018): 639-806.

[3] 方巍, 郑玉, 徐江. 大数据: 概念、技术及应用研究综述. 南京信息工程大学学报, 2014, 6(5): 405-419.

[4] 常杪, 冯雁, 郭培坤, 解惠婷, 王世汶. 环境大数据概念、特征及在环境管理中的应用. 中国环境管理, 2015, 7(6): 26-30.

[5] 赵国栋, 易欢欢, 糜万军, 鄂维南. 大数据时代的历史机遇——产业变革与数据科学. 北京: 清华大学出版社, 2013.

[6] NIMET, Nigeria climate review bulletin (2011). http://www. nimetng. org/uploads/publication/NIMET% 20climate% 20review% 20PDF% 202011%20(1).pdf

[7] 徐子沛. 大数据. 桂林: 广西师范大学出版社, 2012.

[8] 程春明, 李蔚, 宋旭. 生态环境大数据建设的思考. 中国环境管理, 2015, 7(6): 9-13.
[9] 傅伯杰, 刘宇. 国际生态系统观测研究计划及启示. 地理科学进展, 2014, 33(7): 893-902.
[10] 傅伯杰, 牛栋, 于贵瑞. 生态系统观测研究网络在地球系统科学中的作用. 地理科学进展, 2007, 26(1): 1-16.
[11] 戴小廷. 近二十年来生物多样性信息系统的研究进展. 信息技术, 2012, (6): 55-59.
[12] 詹志明, 尹文君. 环保大数据及其在环境污染防治管理创新中的应用. 环境保护, 2016, 44(6): 44-48.
[13] 环境保护部办公厅. 关于印发《生态环境大数据建设总体方案》的通知. (2016-03-08)[2016-03-14]. http://www.mep.gov.cn/gkml/hbb/bgt/201603/t20160311_332712.htm.
[14] 陶雪娇, 胡晓峰, 刘洋. 大数据研究综述. 系统仿真学报, 2013, 25(S1): 142-146.
[15] Wugansha. 大数据漫谈之四: Velocity--天下武功, 唯快不破. (2013-05-28)[2013-05-29]. http://www.huxiu.com/article/15106/1.html.
[16] 孙忠富, 杜克明, 郑飞翔, 尹首一. 大数据在智慧农业中研究与应用展望. 中国农业科技导报, 2013, 15(6): 63-71.
[17] 国家林业局. 中国林业大数据中心已跃升至 PB 级. (2016-02-19)[2016-02-22]. http://www.forestry.gov.cn/main/195/content-844759.html.
[18] 中国存储网. 交通大数据时代需解决的问题分析. (2014-05-14)[2014-05-23]. http://www.chinastor.com/a/dashuju/05143a22014.html.
[19] 中国气象报. 气象大数据时代真的到了吗? (2014-05-06)[2014-05-07]. http://www.cma.gov.cn/2011xwzx/2011xqxxw/2011xqxyw/201405/t20140506_245247.html.
[20] 吴班, 程春明. 生态环境大数据应用探析. 环境保护, 2016, 44(3): 87-89.
[21] 杨晨曦. 全球环境治理的结构与过程研究[D]. 长春: 吉林大学, 2013.
[22] 茅铭晨, 黄金印. 环境污染与公共服务对健康支出的影响——基于中国省际面板数据的门槛分析. 财经论丛: 浙江财经学院学报, 2016, (1): 97-104.
[23] Cattell R. Scalable SQL and NoSQL data stores. SIGMOD Reccrd, 2011, 39(10): 12-27.
[24] Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. Communications of the ACM, 2008, 51: 107-113.
[25] Meijer E. The world according to LINQ. Communications of the ACM, 2011, 54(10): 45-51.
[26] Ghemawat S, Gobio H, Leung S T. The Google file system//Proceedings of the 19th ACM Symposium on Operating Systems Principles. New York, NY, USA: Bolton Landing, 2003, 29-43.
[27] Hey T, Tansley S, Tolle K. The Fourth Paradigm: Data-Intensive Scientific Discovery.Bangalore:Microsoft Research,2009.
[28] Neumeyer L, Robbins B, Kesari A, Nair A. S4: distributed stream computing platform//Proceedings of 2010 IEEE International Conference on Data Mining Workshops. Sydney: IEEE, 2010: 170-177.
[29] 李学龙, 龚海刚. 大数据系统综述. 中国科学: 信息科学, 2015, 45(1): 1-44.
[30] 程学旗, 靳小龙, 王元卓, 郭嘉丰,张铁赢,李国杰. 大数据系统和分析技术综述. 软件学报, 2014, 25(9): 1889-1908.
[31] 洪国伟. 论生物多样性减少的原因及其保护策略. 安徽农学通报, 2010, 16(2): 47-49.
[32] 石虹. 浅谈全球水资源态势和中国水资源环境问题. 水土保持研究, 2002, 9(1): 145-150.
[33] 骆永明. 中国土壤环境污染态势及预防、控制和修复策略. 环境污染与防治, 2009, 31(12): 27-31.
[34] 包晓斌. 防治生态系统退化的对策研究.环境保护, 2012, (20): 48-50.
[35] Zemp M, Frey H, Gärtner-Roer I, Nussbaumer S U, Hoelzle M, Paul F, Haeberli W, Denzinger F, Ahlstrøm A P, Anderson B, Bajracharya S, Baroni C, Braun L N, Cáceres B E, Casassa G, Cobos G, Dávila L R, Delgado G H, Demuth M N, Espizua L, Fischer A, Fujita K, Gadek B, Ghazanfar A, Hagen J O, Holmlund P, Karimi N, Li ZQ, Pelto M, Pitte P, Popovnin V V, Portocarrero C A, Prinz R, Sangewar C V, Severskiy I, Sigurðsson O, Soruco A, Usubaliev R, Vincent C. Historically unprecedented global glacier decline in the early 21st century. Journal of Glaciology, 2015, 61(228): 745-762.
[36] Hinton G, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets. Neural Computation, 2006, 18(7): 1527-1554.
[37] Bengio Y, Lamblin P, Popovici D, Larochelle H, Montreal U. Greedy layer-wise training of deep networks//Plat JC, Koller D, Singer Y,Roweis S T, Eds. Advances in Neural Information Processing Systems19:Proceedings of the 2006 Conference. Cambridge: MIT Press,2007, 19: 153-160.
[38] Dahl GE, Yu D, Deng L, Acero A. Context-Dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Transactionson Audio, Speech, and Language Processing, 2012, 20(1): 30-42.
[39] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks//Pereira F, Burges C J C, Bottou L, Weinberger K Q, eds. Advances in Neural Information Processing Systems 25e. Cambridge: MIT Press, 2012:1097-1105.
[40] 刘智慧, 张泉灵. 大数据技术研究综述. 浙江大学学报: 工学版, 2014, 48(6): 957-972.
[41] Hinton G E. Learning multiple layers of representation. Trends in Cognitive Sciences, 2007, 11(10): 428-434.
[42] Baah G K, Gray A, Harrold M J. On-line anomaly detection of deployed software: a statistical machine learning approach//Proceedings of the 3rd International Workshop on Software Quality Assurance.Portland: ACM, 2006:70-77.
[43] Moeng M, Melhem R. Applying statistical machine learning to multicore voltage & frequency scaling//Proceedings of the 7th ACM International Conference on Computing Frontiers.Bertinoro:ACM, 2010: 277-286.
[44] 陈左宁, 王广益, 胡苏太, 韦海亮. 大数据安全与自主可控. 科学通报, 2015, 60(5/6): 427-432.
[45] 赵岑, 李梦然, 金日峰. 大数据时代关于隐私的思考. 科学通报, 2015, 60(5/6): 450-452.
[46] 杨曦, Gul J, 罗平. 云时代下的大数据安全技术. 中兴通讯技术, 2016, 22(1): 14-18.
[47] 王世伟. 论大数据时代信息安全的新特点与新要求. 图书情报工作, 2016, 60(6): 5-14.
[48] 冯伟. 大数据时代面临的信息安全机遇和挑战. 中国科技投资, 2012, (34): 49-53.